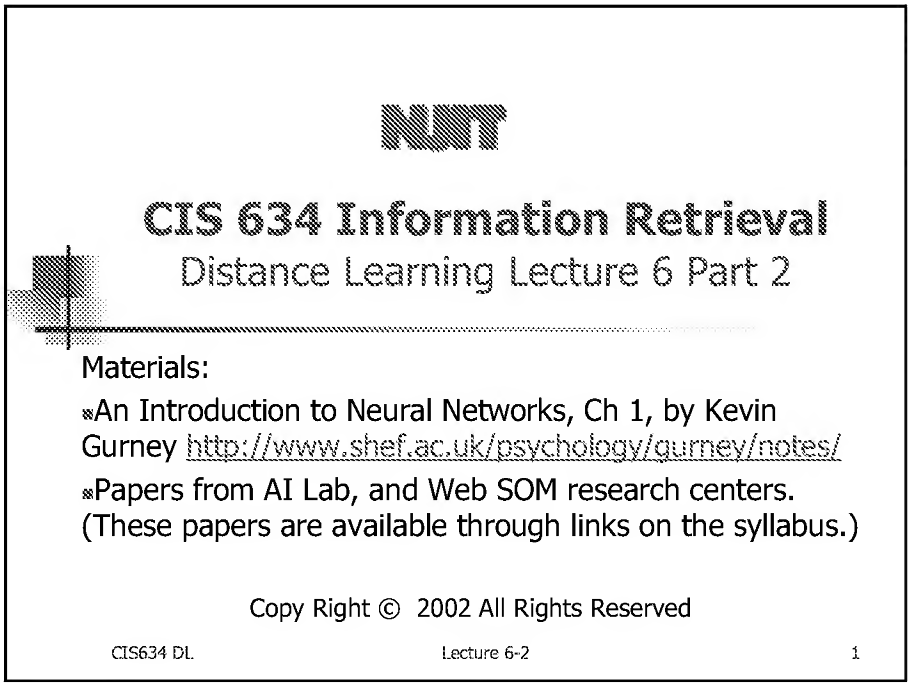


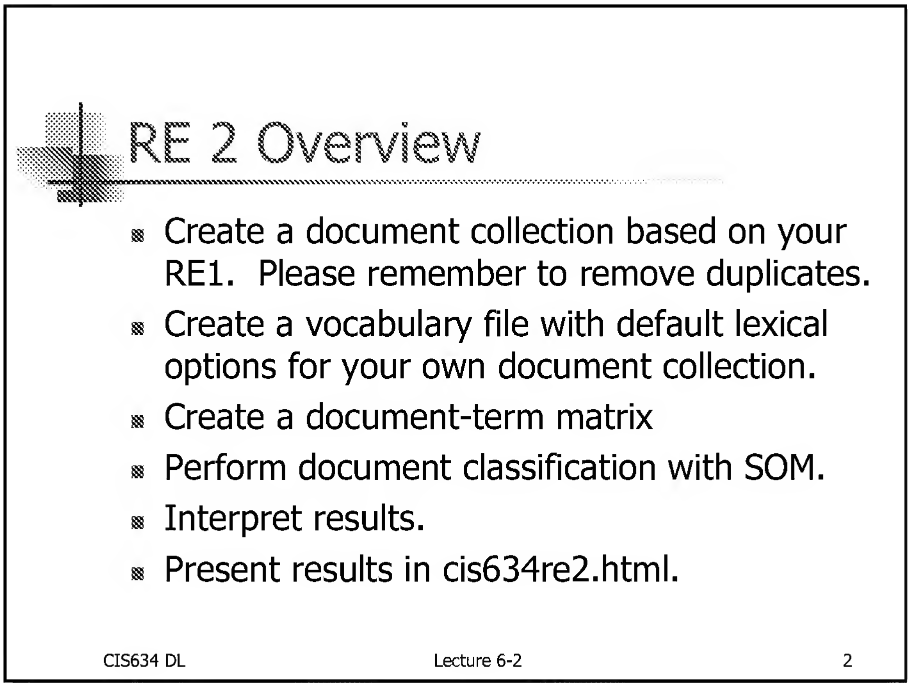
NJIT

CIS 634 Information Retrieval

Distance Learning Lecture 6 Part 2

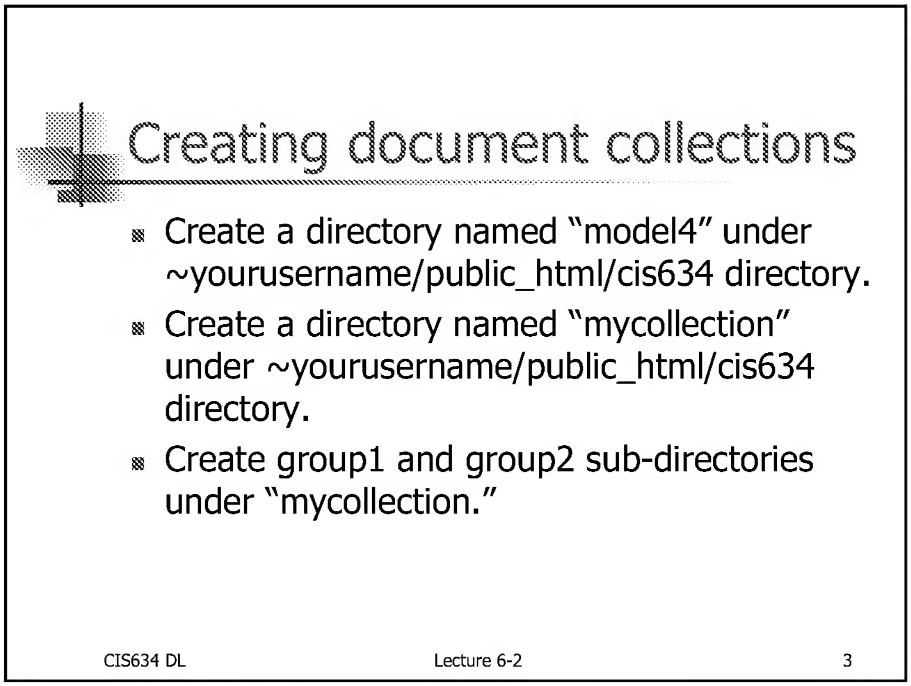
Materials:

- An Introduction to Neural Networks, Ch 1, by Kevin Gurney http://www.shef.ac.uk/psychology/gurney/notes/
- Papers from AI Lab, and Web SOM research centers. (These papers are available through links on the syllabus.)



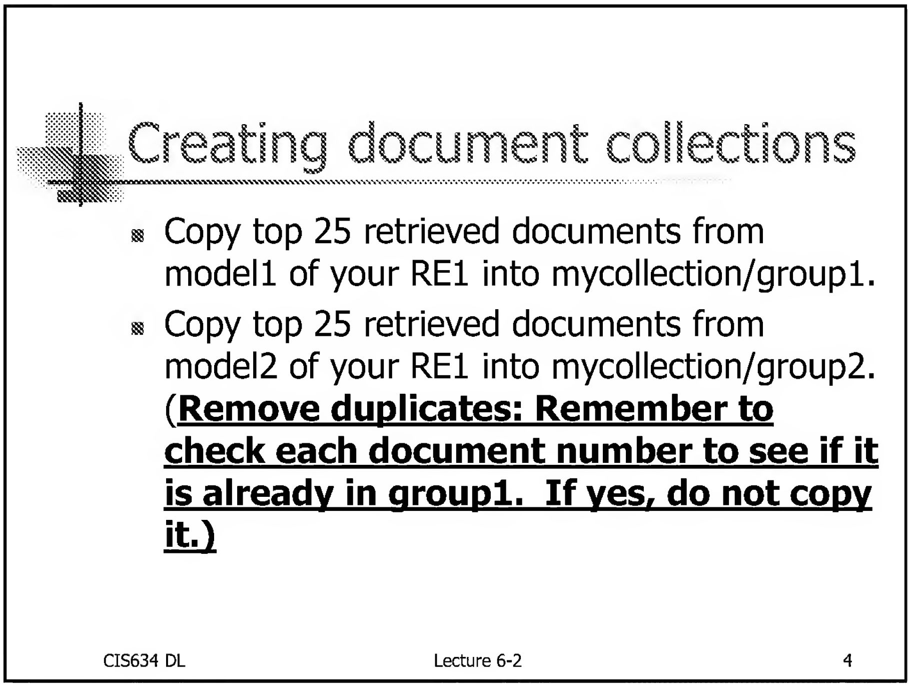
RE 2 Overview

- Create a document collection based on your RE1. Please remember to remove duplicates.
- Create a vocabulary file with default lexical options for your own document collection.
- Create a document-term matrix
- Perform document classification with SOM.
- Interpret results.
- Present results in cis634re2.html.



Creating document collections

- Create a directory named "model4" under ~yourusername/public_html/cis634 directory.
- Create a directory named "mycollection" under ~yourusername/public_html/cis634 directory.
- Create group1 and group2 sub-directories under "mycollection."



Creating document collections

- Copy top 25 retrieved documents from model1 of your RE1 into mycollection/group1.
- Copy top 25 retrieved documents from model2 of your RE1 into mycollection/group2. (**Remove duplicates: Remember to check each document number to see if it is already in group1. If yes, do not copy it.)**

Creating document collections

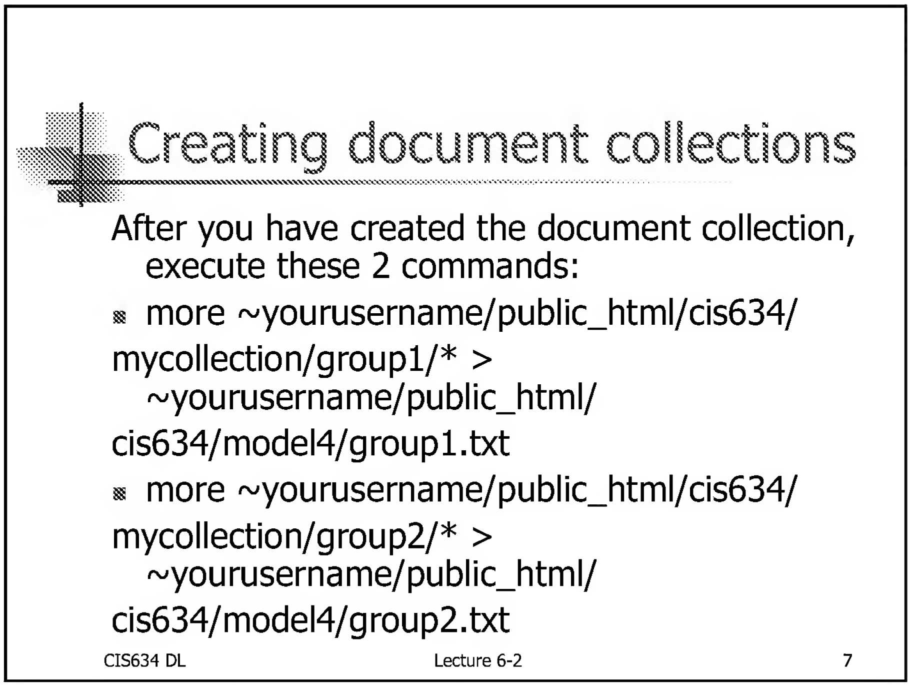
- Copying retrieved documents into mycollection/group1:
  - Make sure you are in ~yourusername/public_html/cis634/mycollection/group1

Creating document collections (cont)

- At the system prompt, type in: cp filename .
- . (the dot sign) means the current directory
- ex: cp /afs/cad/u/w/u/wu/cis634

/tc/lisa/text/group0/doc_306 .

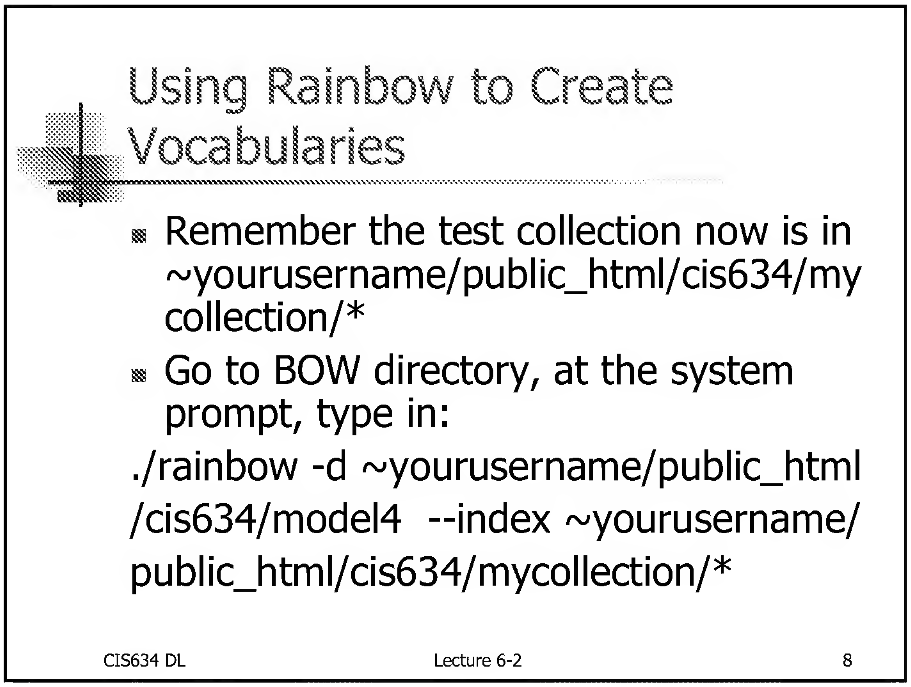
- Repeat the same process for group2, which contains retrieved documents from model2 of RE1 (remember to remove duplicates).



Creating document collections

After you have created the document collection, execute these 2 commands:

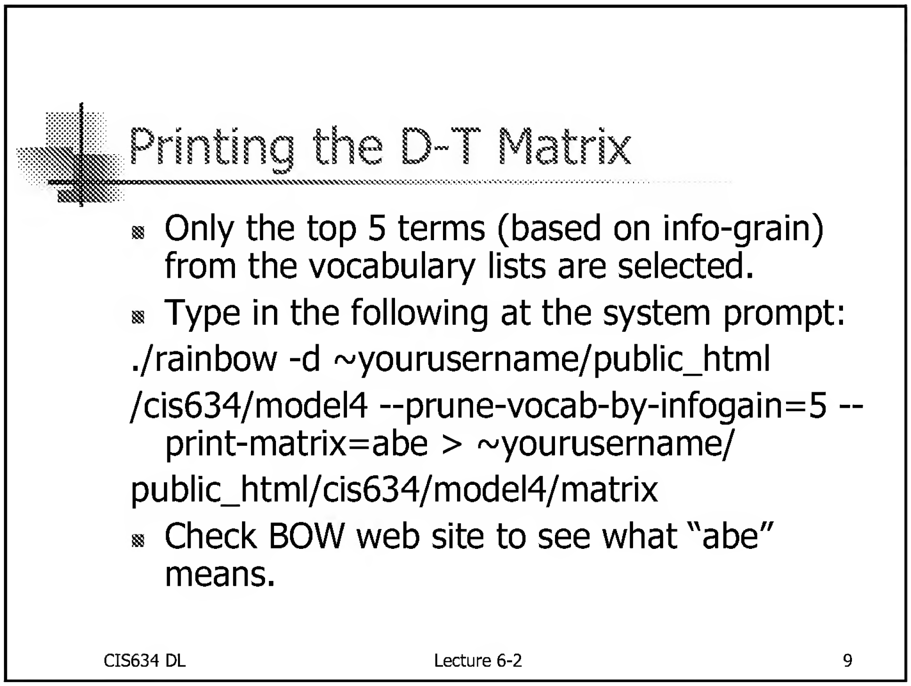
- more ~yourusername/public_html/cis634/
mycollection/group1/* >
~yourusername/public_html/
cis634/model4/group1.txt
- more ~yourusername/public_html/cis634/
mycollection/group2/* >
~yourusername/public_html/
cis634/model4/group2.txt



Using Rainbow to Create Vocabularies

- Remember the test collection now is in ~yourusername/public_html/cis634/mycollection/*
- Go to BOW directory, at the system prompt, type in:

./rainbow -d ~yourusername/public_html
/cis634/model4 --index ~yourusername/
public_html/cis634/mycollection/*

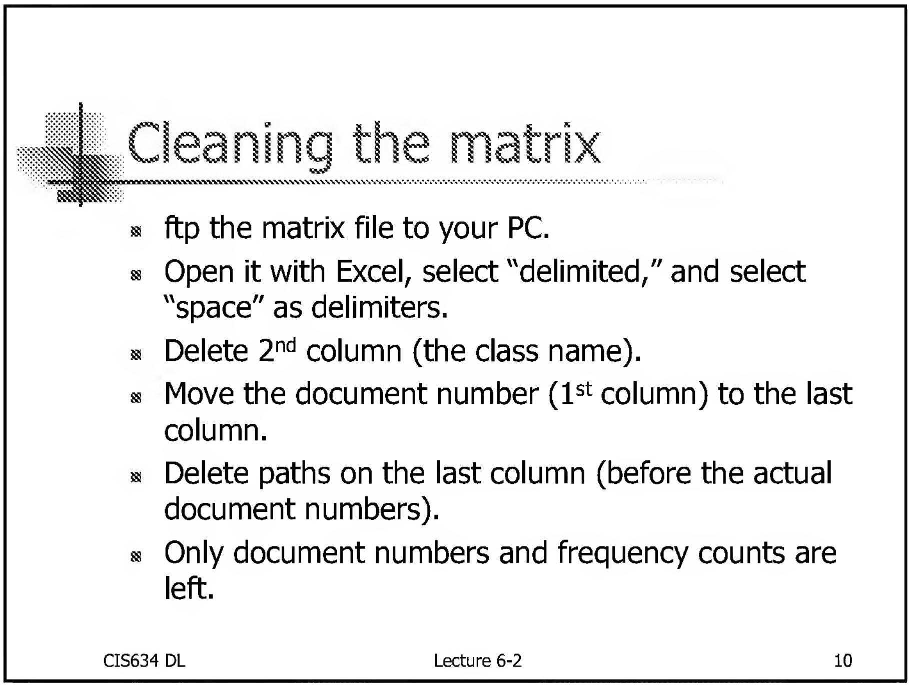


Printing the D-T Matrix

- Only the top 5 terms (based on info-grain) from the vocabulary lists are selected.
- Type in the following at the system prompt:

```
./rainbow -d ~yourusername/public_html
/cis634/model4 --prune-vocab-by-infogain=5 --
   print-matrix=abe > ~yourusername/
public_html/cis634/model4/matrix
```

- Check BOW web site to see what "abe" means.



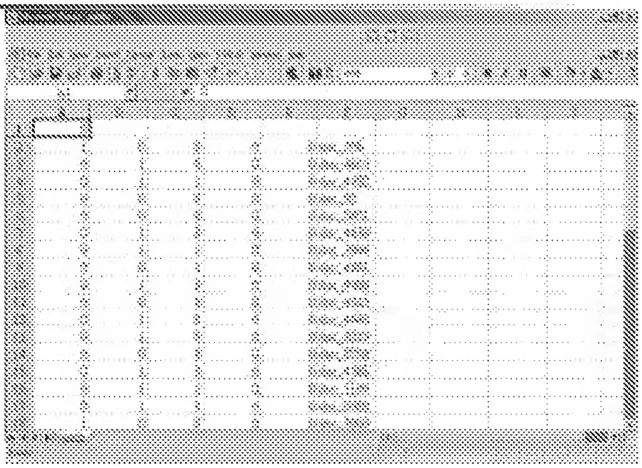
Cleaning the matrix

- ftp the matrix file to your PC.
- Open it with Excel, select "delimited," and select "space" as delimiters.
- Delete 2nd column (the class name).
- Move the document number (1st column) to the last column.
- Delete paths on the last column (before the actual document numbers).
- Only document numbers and frequency counts are left.

Cleaning the matrix

- Insert an empty row before the first row. Type in 5 (for 5 properties) in the very first cell.
- Save the file in "Text" (Tab delimited) format, the file name is matrix.txt
- Upload this file back to model4 directory.
- ****However, Nenet uses *.dat for input matrix files. You will have to specify the file type as "all files," when opening data file in Nenet.

The final matrix (to be saved in Text (Tab delimited) format)



SOM toolkit for Windows -- Nenet

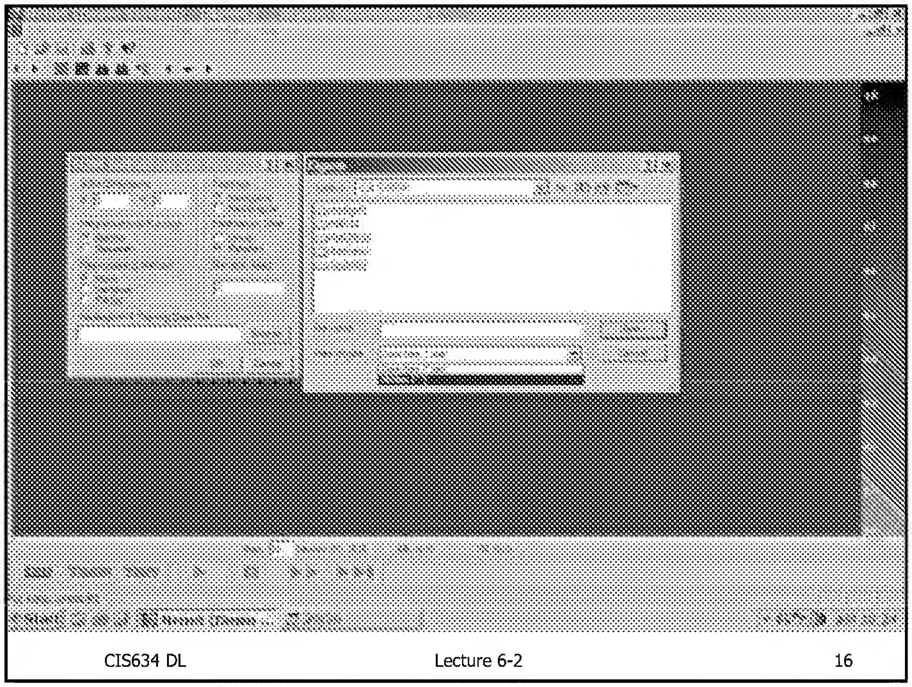
- The trial version is available at: http://koti.mbnet.fi/~phodju/nenet/Nenet/General.html
- Trial version has limited capability: up to 8 properties, 6x6 dimensions (36 neurons). However, if the matrix has 8 properties, Nenet seems to have trouble with it. So, please limit your raw data (matrix and matrix.txt files) to exact 5 properties.

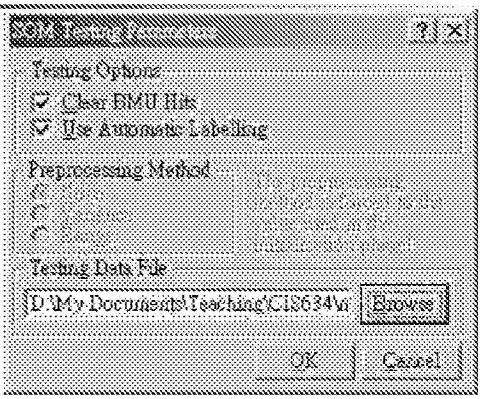
Download and Install Nenet

- Create a file folder named temp. Download all three zip files to temp and uncompress them with WinZip. Install the software by clicking on **setup.exe**.
- If your PC doesn't have WinZip, download it here: http://www.winzip.com/

Nenet Demo & Dataset

- Interactive demo:
  http://koti.mbnet.fi/~phodju/nenet/Nenet/InteractiveDemo.html
- For your RE2, the initial dataset, training dataset, and test dataset are the same, that is "matrix.txt".
- Nenet uses *.dat for input matrix files. You will have to specify the file type as "all files," when opening data file in Nenet.
- Remember to select "Use Automatic Labeling" at the testing stage. (or your map will not have document numbers as labels!!)





Testing Options

☑ Clear BMU Hits

☑ Use Automatic Labelling

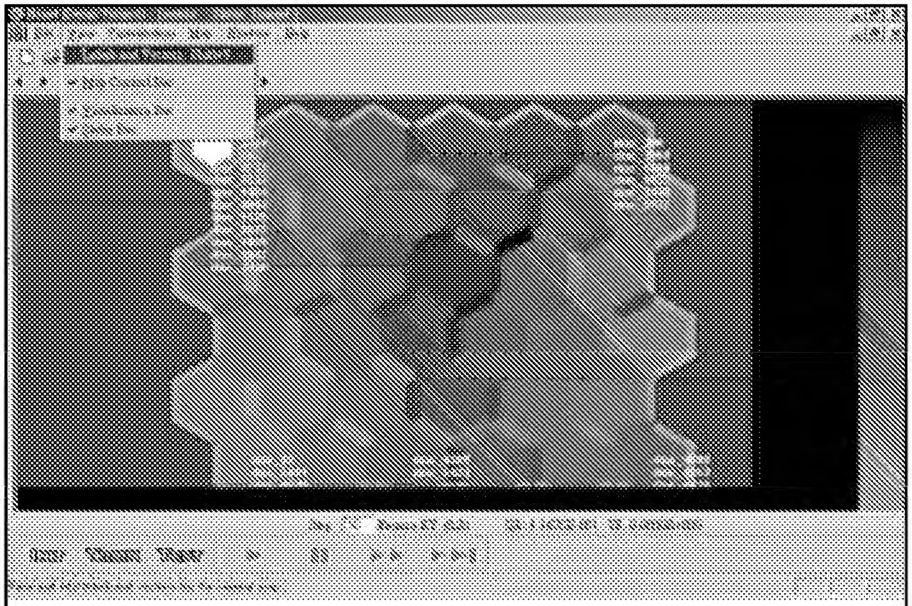
Preprocessing Method

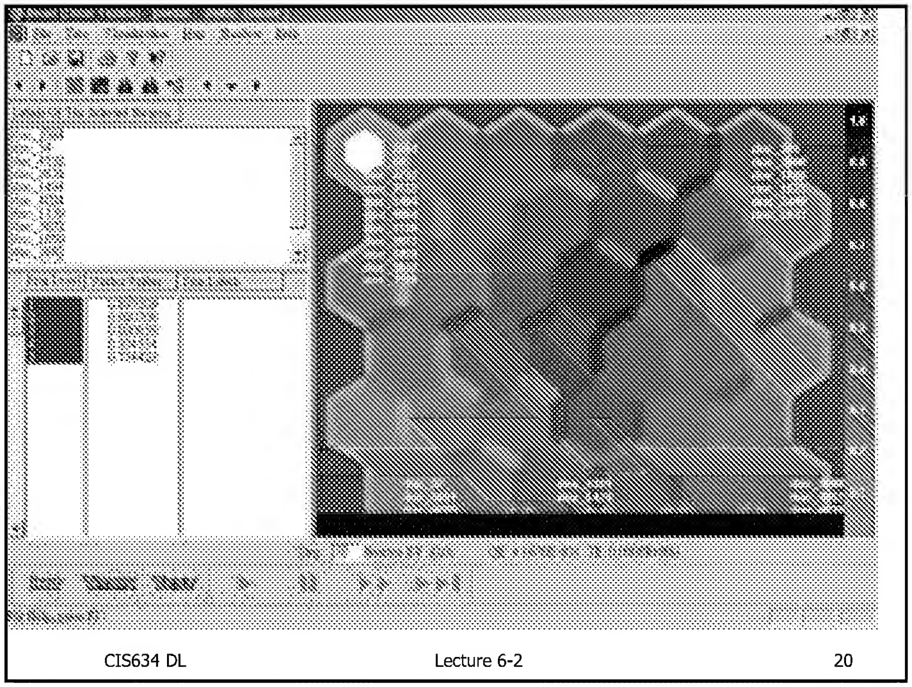
Testing Data File

D:\My Documents\Teaching\CIS634\n

View Results on the Feature Map

- After training and testing, Nenet presents the results on a map similar to slide 19.
- Click on "view" → "labels and vectors," Nenet will bring you to a screen similar to slide 20.
- Click on any neuron on the map that has document numbers on it, you will see a list of document numbers associated with that neuron.

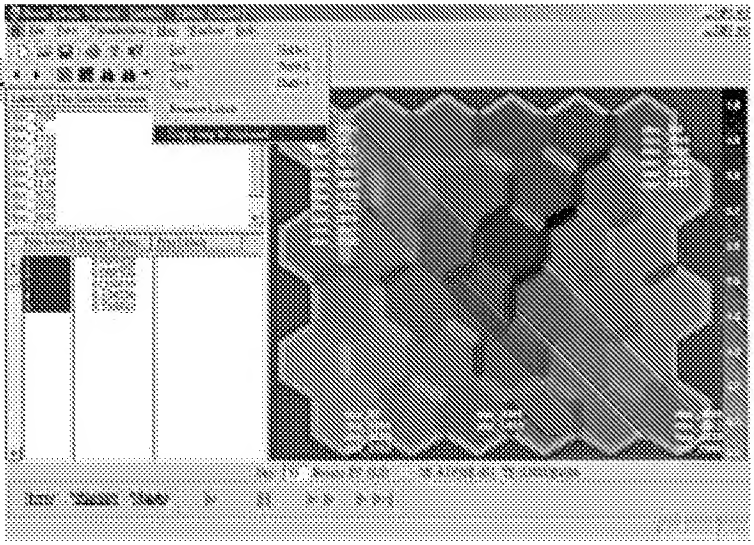




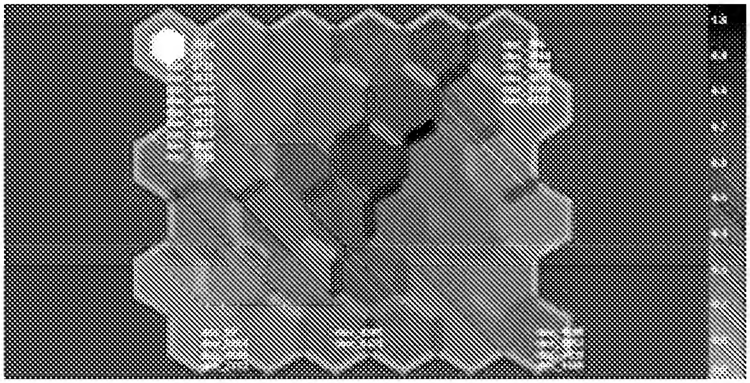
How are Doc# mapped to the neuron?

- When labeling, each document vector is compared to the final vector of weights of each neuron.
- The best matching neuron determines where the document# will be located on the map.

Copy Map to Clipboard



Save this map in matrix.jpg or matrix.gif file, and upload the map to model4 directory

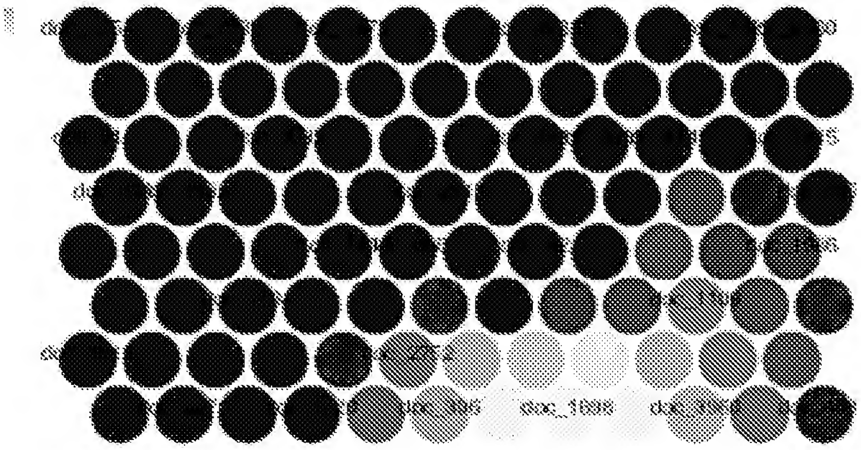


Your tasks

- Use the D-T matrix (**matrix.txt**) created earlier for document clustering with Nenet.
- Follow the instructions on the interactive demo.
- Save the final results in **matrix.cod** file.
- Upload the **matrix.txt**, **matrix.cod,** and **matrix.jpg** to model4 directory.
- Create RE2 page, format: http://web.njit.edu/~wu/cis634/cis634re2.html

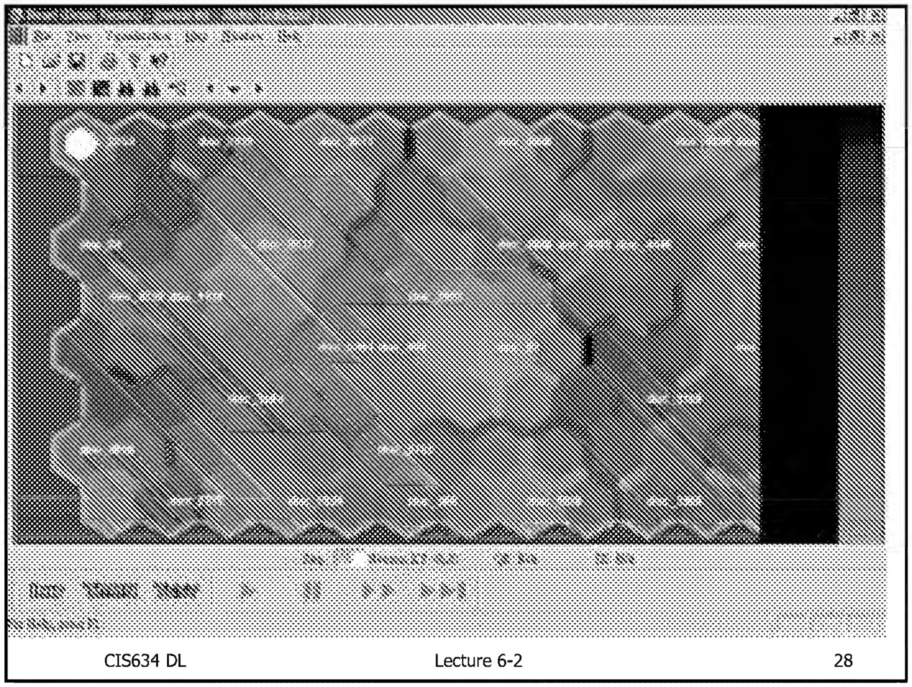
An Alternative

- Nenet trial version has limited capabilities: up to 8 properties, and 2000 records.
- An alternative: SOM_PAK does not have restrictions on the size of datasets. The original D-T matrix for the output map in slide 26 has 25 properties (terms).
- SOM_PAK is located at: ~wu/IR_Tools/som
- However, the postscript map files created by SOM_PAK are hard to read.

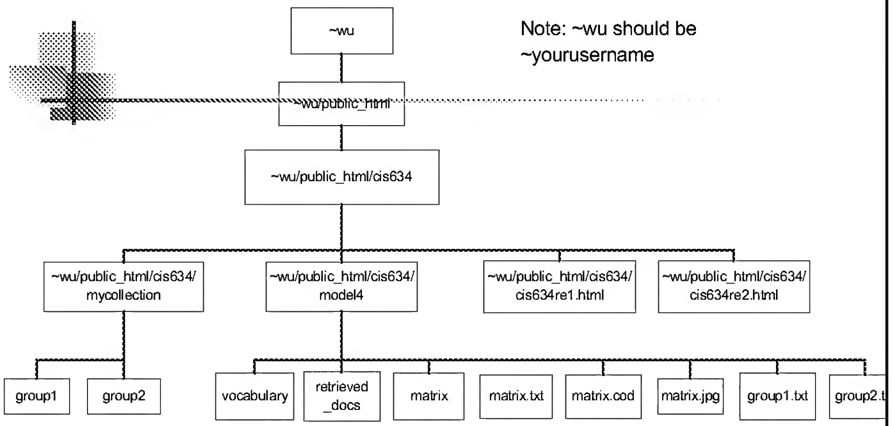


An Alternative to the Alternative

- Use SOM_PAK for the whole process and create matrix.cod output map file. (Save the file in your model4 directory.)
- Download the matrix.cod to your PC, and read the file with Nenet. (file →open→ matrix.cod)
- No instructions on SOM_PAK will be provided. You will have to read SOM_PAK manual by yourself.
- However, those use SOM_PAK to process the D-T matrix with higher number of properties, will receive 2 extra points.



Directory structure of your UNIX account should look like this

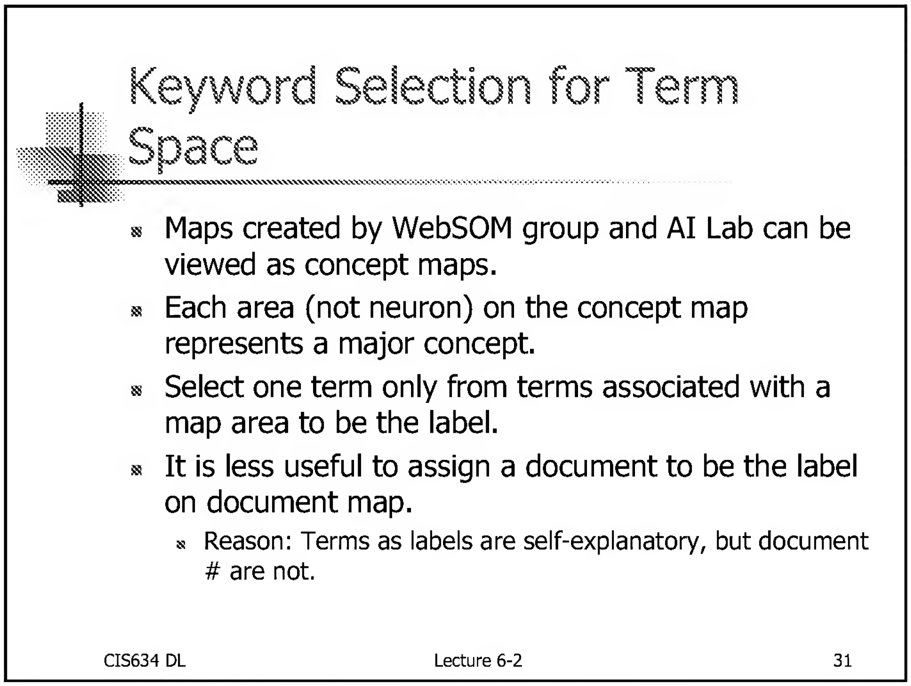

~wu
~wu/public_html
~wu/public_html/cis634
~wu/public_html/cis634/ mycollection
~wu/public_html/cis634/ model4
~wu/public_html/cis634/ cis634re1.html
~wu/public_html/cis634/ cis634re2.html
group1
group2
vocabulary
retrieved _docs
matrix
matrix.txt
matrix.cod
matrix.jpg
group1.txt
group2.t

Note: ~wu should be ~yourusername

(mycollection has two sub-directories. group1 stores retrieved documents from model1 of RE1, and group2 stores those from model2)

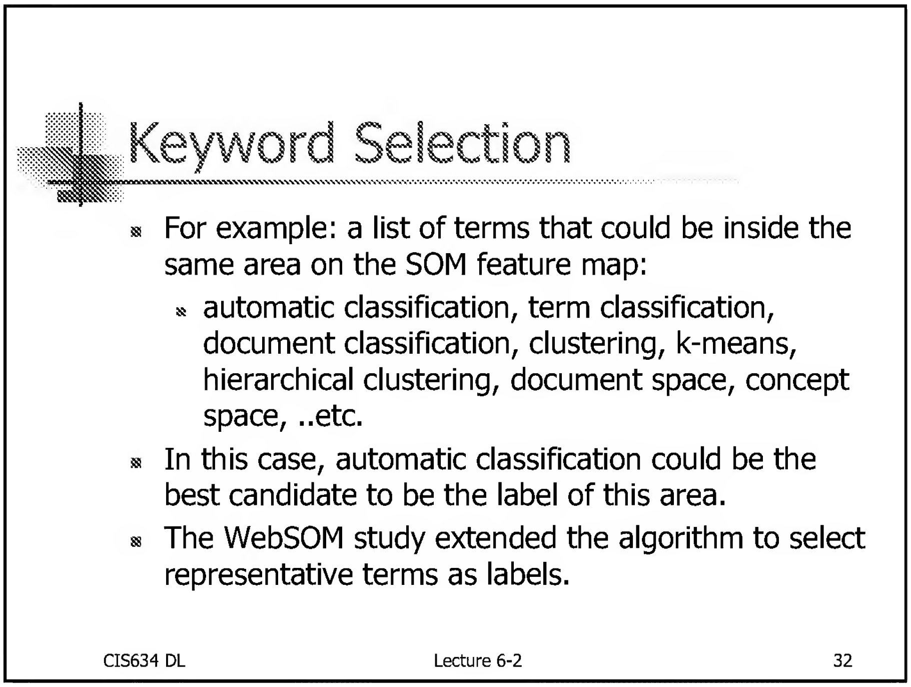
What are the differences between the results from RE2 and WebSOM

- RE2: a neuron can have multiple document # associated with it, namely, many labels.
- WebSOM: each area is labeled with one term only.
  - Note: When talking about term space, researchers tend to use "terms" and "concepts" interchangablely.
- What makes them different?



Keyword Selection for Term Space

- Maps created by WebSOM group and AI Lab can be viewed as concept maps.
- Each area (not neuron) on the concept map represents a major concept.
- Select one term only from terms associated with a map area to be the label.
- It is less useful to assign a document to be the label on document map.
  - Reason: Terms as labels are self-explanatory, but document # are not.



Keyword Selection

- For example: a list of terms that could be inside the same area on the SOM feature map:
  - automatic classification, term classification, document classification, clustering, k-means, hierarchical clustering, document space, concept space, ..etc.
- In this case, automatic classification could be the best candidate to be the label of this area.
- The WebSOM study extended the algorithm to select representative terms as labels.

How to Use SOM for 2nd Type Classification?

- Initialization and training process is the same.
- The only difference is the testing part – use a different set of D-T matrix.
- How can the resultant maps be used?
  - Automatic Cataloging